Data for: A Literature Review on Methods for the Extraction of Usage Statements of Software and Data

Frank Krüger David Schindler

September 12, 2019

Software and data have become major components of modern research [1], which is also reflected by an increased number of software usages. Knowledge about used software and data would provide researchers a better understanding of the results of a scientific investigation and thus foster it's reproducibility. Software and data are, however, often not formally cited but their usage is mentioned in the main text. In order to assess the state of the art in extraction of such usage statements, we performed a literature review. We provide an overview of existing methods for the identification of usage statements of software and data in scientific articles. This analysis mainly focuses on technical approaches, the employed corpora, and the purpose of the investigation itself. We found four different classes of approaches that are used in the literature: 1.) term search, 2.) manual extraction, 3.) rule-based extraction, and 4.) extraction based on supervised learning.

1 General Information

Affiliations All authors are with the Institute of Communications Engineering, University of Rostock, Rostock, Germany

E-mail Frank.Krueger@uni-rostock.de

Keywords software and data citation, named entity recognition, literature review.

Language English

2 Description

The dataset contains all supplements for a literature review on methods for the extraction of usage statements of data and software from scientific literature. The following files are included:

01_table_of_results.pdf : Table of classification including list of references

02_retrieval_protocol.pdf : Detailed description of all steps of the literature search and selection

03_first_13_articles.pdf : Initial set of articles that were retrieved by unstructured literature search

04_applied_queries.pdf : Database queries that were used to retrieve the actual literature from the different literature databases

05_database_query_result-1392.csv : Result set from applying the queries to the literature databases

06_title_reduced_set-244.csv : Result set that was created by manually scanning through the titles of the initial result set

07_abstract_reduced_set-46.csv : Result set that was created by manually scanning through the abstracts of the reduced result set

08_query_scripts.zip : Scripts that were used for automatic querying of the literature databases

09_supplement_code.zip : Code to implement the neural network for named entitiy recognition of databases and software